

# Neurowiki

**Vulcan Inc.**

*in conjunction with*

**Allen Institute for Brain Science**

# Today we are Discussing...

- What is the use case and who requested it?
- How does this apply to a Semantic Media Wiki?
- How do you import and normalize thousands of triples worth of gene RDF triples?
- Creating instance pages without knowing exactly what will be displayed on them.
- Embedding SPARQL within SMW Templates.
- Expanding existing instance pages with Semantic Results Formatters, a PHP view layer, and creating dynamic charts.

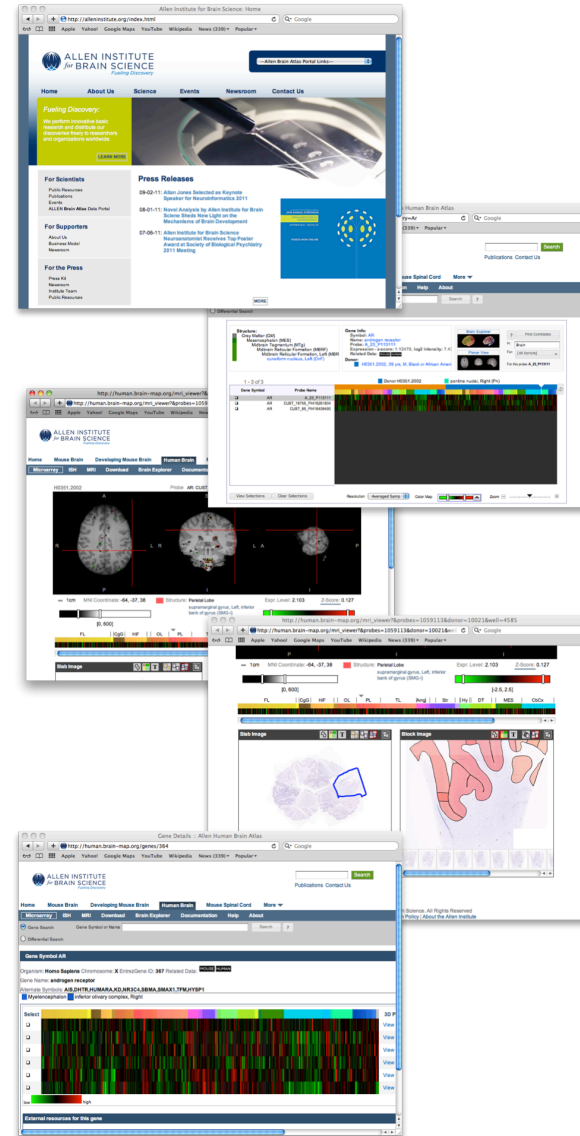
# What is the Allen Institute?

- Launched in 2003 with seed funding from founder and philanthropist Paul G. Allen.
- Serving the scientific community is at the center of our mission to accelerate progress toward understanding the brain and neurological systems.
- The Allen Institute's multidisciplinary staff includes neuroscientists, molecular biologists, informaticists, and engineers.

“The Allen Institute for Brain Science is an independent 501(c)(3) nonprofit medical research organization dedicated to accelerating the understanding of how the human brain works.”

# Human Brain Map

- Open, public online access
- A detailed, interactive three-dimensional anatomic atlas of the "normal" human brain
- Data from multiple human brains
- Genomic analysis of every brain structure, providing a quantitative inventory of which genes are turned on where
- High-resolution atlases of key brain structures, pinpointing where selected genes are expressed down to the cellular level
- Navigation and analysis tools for accessing and mining the data



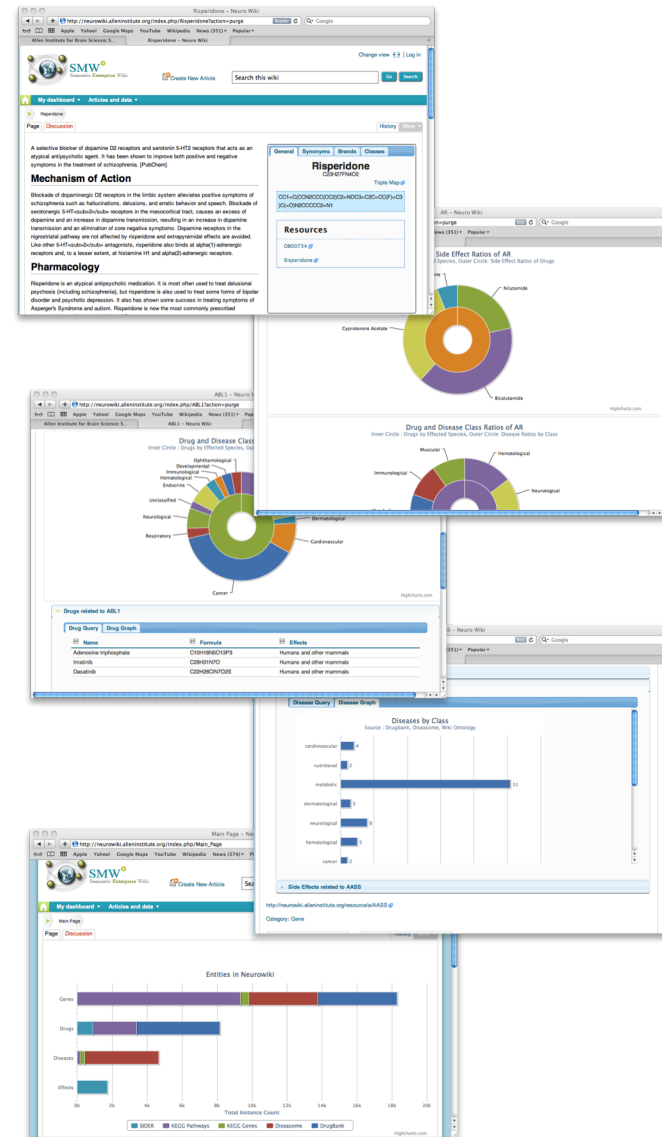


# What is Neurowiki?

- A joint project between Vulcan Inc. and the Allen Institute to build a Semantic Wiki mapping genetic instances.
- A finished prototype testing the import pipelines and display components for combining 5 major RDF datasets from 4 different sources.
- Current planning includes mapping complete datasets, curating a better ontology, creating multiple ontology management for a user class, and importing scientific papers.

# Biological Linked Data Map

- Open, public online access
- Data from multiple RDF data stores
- Complete import pipeline using LDIF framework
- Outlines of each imported instance embedding inline wiki properties and providing views of imported properties from original RDF datasets
- Charting tools that 'pivot' SPARQL queries providing several views of each query
- Navigation and composition tools for accessing and mining the data



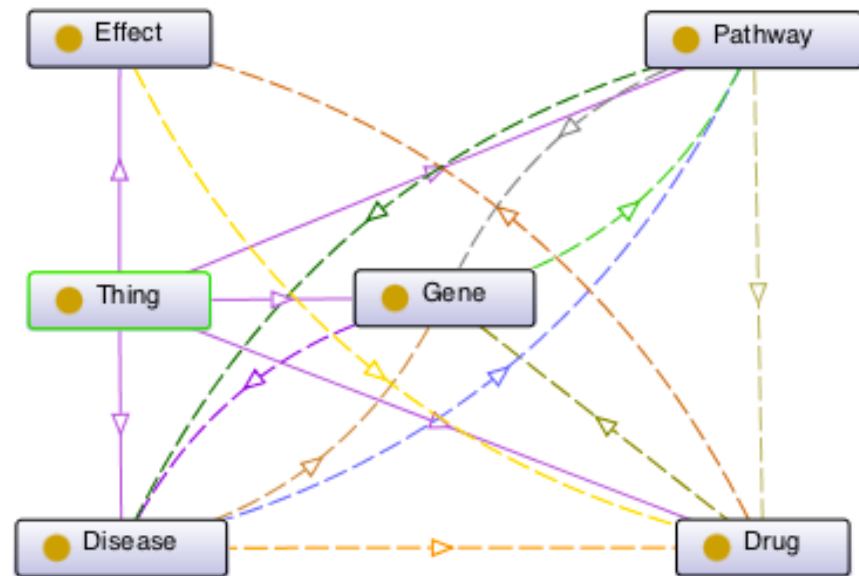
# Where did we get the data?

- **KEGG : Kyoto Encyclopedia of Genes and Genomes**
  - “**KEGG GENES** is a collection of gene catalogs for all complete genomes generated from publicly available resources, mostly NCBI RefSeq.”
- **Diseasome**
  - “The **Diseasome** website is a disease/disorder relationships explorer and a sample of an innovative map-oriented scientific work. Built by a team of researchers and engineers, it uses the Human Disease Network dataset.”
- **DrugBank**
  - “The **DrugBank** database is a unique bioinformatics and cheminformatics resource that combines detailed drug data with comprehensive drug target information.”
- **SIDER**
  - “**SIDER** contains information on marketed medicines and their recorded adverse drug reactions. The information is extracted from public documents and package inserts.”

# Wiki Ontology Map

- Genes
  - DrugBank : 4,553
  - Disasome : 3,919
  - KEGG : 9,841
- Diseases
  - Disasome : 4,213
  - KEGG : 459
- Drugs
  - DrugBank : 4,772
  - KEGG : 2,482
  - SIDER : 924
- Effects
  - SIDER : 1,737
- Pathways
  - KEGG : 28,442

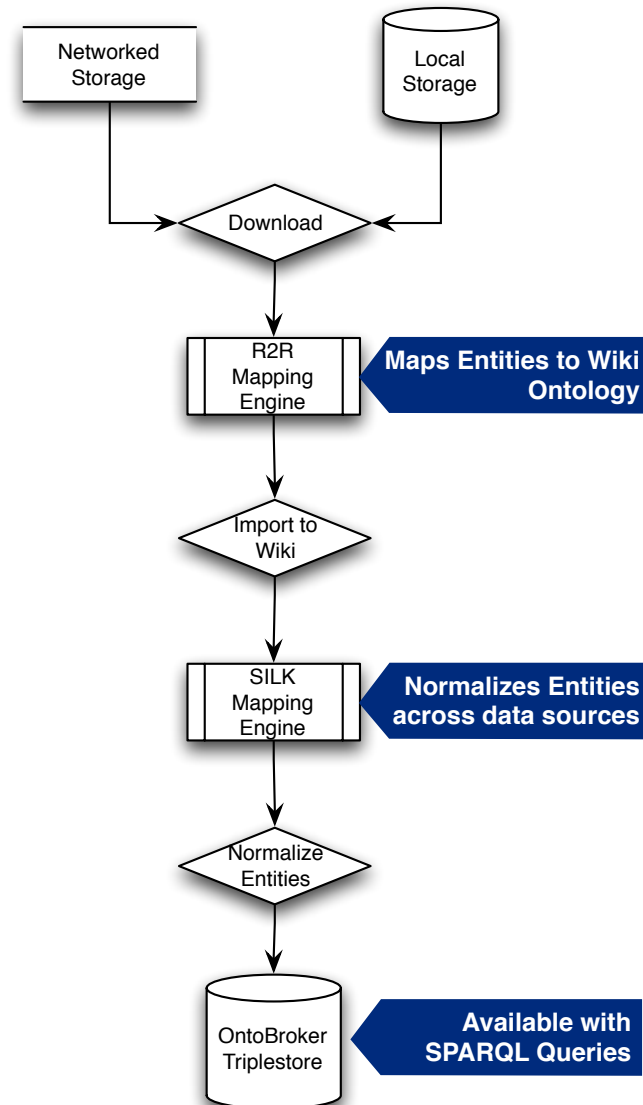
**61,342** Instances Available  
for Import



We chose to intentionally simplify the ontology due to disagreements between researchers about entity relationships and subclasses.

# Importing and Mapping the Linked Data

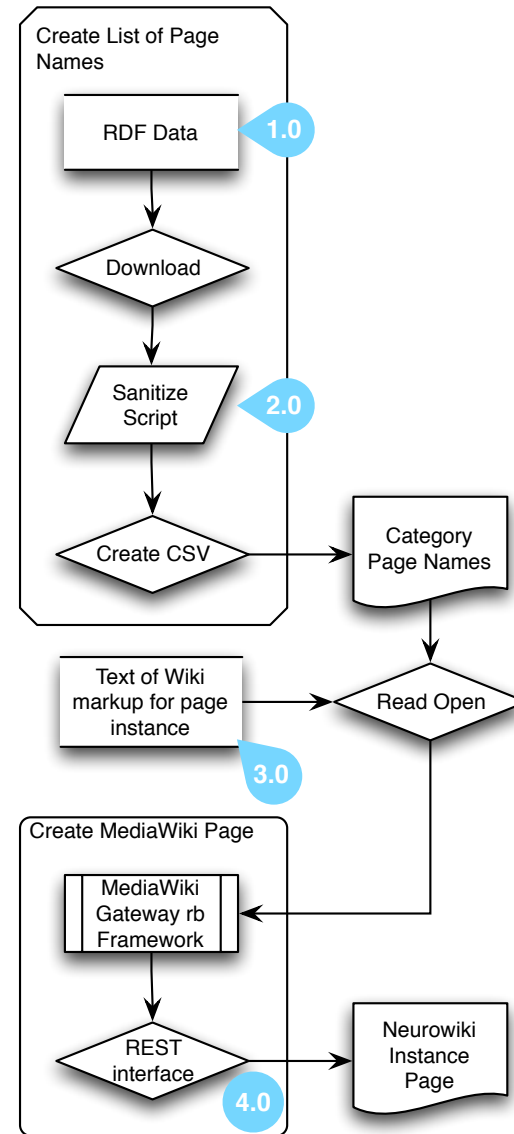
- R2R
  - **32,900** instances were converted to the wiki ontology.
  - **583,746** properties mapped
  - Pathways were ignored for wiki ontology import, but are available within the triple store KEGG Pathway graph.
- SILK
  - **20,849** instances available in wiki ontology after SILK normalization
  - Instance merging effected drugs, genes, and diseases across datasets.
- OntoBroker Triplestore



# Creating Instance Wiki Pages

The triplestore now contained **tens of thousands** of recognized category instances. Creating the pages would require a bot.

1. Fetch the RDF dumps from an active D2R server
2. Use regex to fetch the rdf:label property that was mapped by R2R as an instance name
3. Open category specific text file of wiki markup (page of template includes)
4. Contact Neurowiki and request a new page from the list of names with the category content





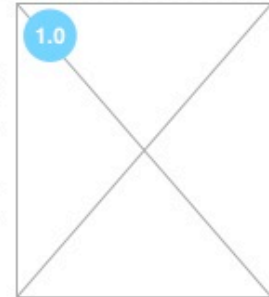
# Four Initial Templates for Each Instance by Category

1. Custom infobox within outline template
  - Visible inline properties
2. Outline template providing instance information
3. Widget template displaying dynamic charts or third party services
  - Donut charts and disease Twitter feed
4. Broad table SPARQL queries showing instance relationships
5. Hidden inline properties for other extensions

## ClassOutline Template

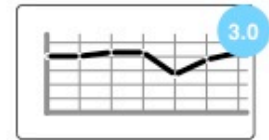
2.0 Lorem ipsum dolor sit amet, consectetur adipiscing elit. Aliquam lacinia, dolor sed condimentum fringilla, augue diam mattis lectus, eget vestibulum elit nulla id arcu. Lorem ipsum dolor sit amet, consectetur adipiscing elit. Vestibulum et sem et lacus lobortis bibendum.

EX: {{{GeneOutline|{{PAGENAME}}}}



## ClassWidgets Template

EX: {{{GeneWidgets|{{PAGENAME}}}}



## DefaultClassQueries Template

EX: {{{DefaultGeneQueries|{{PAGENAME}}}}



## ClassSemantics Template

EX: {{{GeneSemantics|{{PAGENAME}}}}

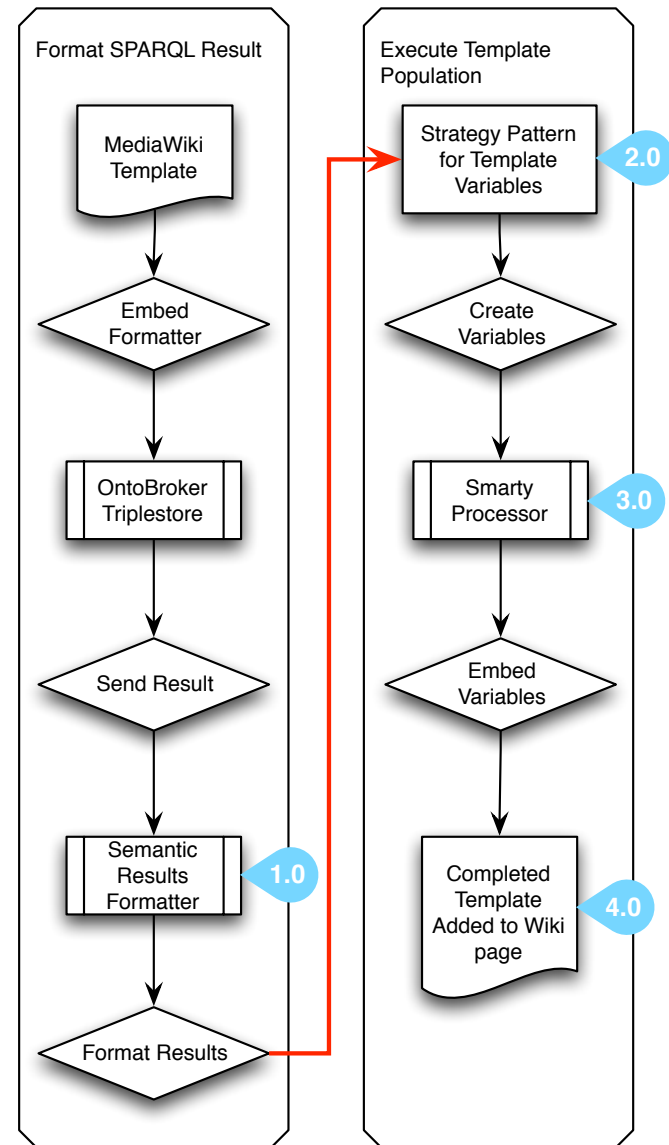
EX: [[Category:Gene]]

5.0



# Adding a True View Layer to Semantic Results Formatter

1. Standard Semantic Results Formatter.
2. Classes made extending the stock formatter and providing a library of strategy classes for rendering different graph types.
3. Assembled template variables are given to the Smarty PHP templating engine with a path to the template file.
4. Template is populated and inserted / injected into the complete MediaWiki page.



# Embedding SPARQL

## Semantic Results Formatters

- Every piece of content on every instance page is generated by Semantic Result Formatters interpreting SPARQL results.
- Most inline properties are embedded in templates returned by SPARQL formatters.
- All 3 dynamic graph types are interpreting results of SPARQL queries and injecting a JavaScript template into the head of the page.
- The outline template takes selected predicates and objects from a SPARQL query, defined in the query embedding, and generates an HTML template for the page.

```
SELECT DISTINCT ?p ?o ?p1 ?o1 {  
  GRAPH ?G {
```

```
    ?gene rdfs:label "{1}";  
    diseasome:bio2rdfSymbol ?sym ;  
    ?p ?o ;  
    .
```

Selects a gene  
from a set of  
**3,919** in  
Diseasome

```
  }  
OPTIONAL {  
  GRAPH ?G1 {
```

```
    ?gene2 drugbank:bio2rdfSymbol ?sym ;  
    ?p1 ?o1 ;
```

Attempts to find  
much more  
information about  
gene using  
bio2rdfSymbol

```
  }
```

```
SELECT DISTINCT ?group1 ?item1 ?group2 ?item2 {  
  GRAPH ?G {
```

```
    ?target drugbank:geneName "{{{1}}}" ;  
    drugbank:geneName ?geneName ;
```

```
    .
```

```
    ?drug drugbank:target ?target ;  
    drugbank:genericName ?item2 ;  
    drugbank:affectedOrganism ?group2 ;
```

```
    .
```

Find the inner  
ring and group by  
organisms  
affected by drugs  
targeting this gene

```
  }  
  GRAPH ?G1 {
```

```
    ?siderDrug sider:drugName ?item2 ;
```

```
    rdfs:label ?group1 ;
```

```
    sider:sideEffect ?effect;
```

```
    .
```

```
    ?effect rdfs:label ?item1 .
```

```
  }
```

```
}
```

In the outer ring  
group the side effects  
by drugs targeting  
the gene used to  
form the inner ring

# Final Application Stack

## Frameworks

- MediaWiki
- LDIF
- Semantic MediaWiki
- SMW+

## Extensions

- Data Import
- Semantic Results Format
- SMWHalo
- Enhanced Retrieval
- WikiTags

## View Libraries

- JQuery
- JQueryUI
- Smarty
- Highcharts

# Bugs, Bottlenecks, and Unexpected Features

- In original data `rdf:label` was not normalized
  - `Breast_cancer_302400`
- Ignored labels that would break MediaWiki URL schemes
  - `4'-acetate(30px-[4l]''')-di_tetraoxil''(40-Cl2)`
- Certain drug compounds are related to almost everything
  - Calcium
  - Vitamin A
- Page create bot had to re-edit pages upon completion to register inline properties with SMW & SMW+ extensions.

# Neurowiki in Action!

- Which drugs are used in Chemotherapy?
- What are the dangers of Propofol?
- How are base entities like Calcium represented?
- How are new inline properties added to entities?
  - Can these be searched?
  - Can these be queried using ASK?
- Do existing extensions work with the framework?



# Demo Links

- [http://neurowiki.alleninstitute.org/index.php/Main\\_Page](http://neurowiki.alleninstitute.org/index.php/Main_Page)
- <http://neurowiki.alleninstitute.org/index.php/AR>
- <http://neurowiki.alleninstitute.org/index.php/ABL1>
- <http://neurowiki.alleninstitute.org/index.php/AASS>
- <http://neurowiki.alleninstitute.org/index.php/AAC2>
- <http://neurowiki.alleninstitute.org/index.php/Thioridazine>
- <http://neurowiki.alleninstitute.org/index.php/Risperidone>
- <http://neurowiki.alleninstitute.org/index.php/Propofol>
- <http://neurowiki.alleninstitute.org/index.php/Calcium>